Unsupervised mining of audiovisually consistent segments in videos with application to structure analysis

Mathieu Ben, Guillaume Gravier

UNSUPERVISED MINING OF AUDIOVISUALLY CONSISTENT SEGMENTS IN VIDEOS WITH APPLICATION TO STRUCTURE ANALYSIS

Mathieu Ben

INRIA Rennes
35042 Rennes Cedex, France
mathieu.ben@inria.fr

Guillaume Gravier

IRISA-CNRS
35042 Rennes Cedex, France
guillaume.gravier@irisa.fr

ABSTRACT

In this paper, a multimodal event mining technique is proposed to discover repeating video segments exhibiting audio and visual consistency in a totally unsupervised manner. The mining strategy first exploits independent audio and visual cluster analysis to provide segments which are consistent in both their visual and audio modalities, thus likely corresponding to a unique underlying event. A subsequent modeling stage using discriminative models enables accurate detection of the underlying event throughout the video. Event mining is applied to unsupervised video structure analysis, using simple heuristics on occurrence patterns of the events discovered to select those relevant to the video structure. Results on TV programs ranging from news to talk shows and games, show that structurally relevant events are discovered with precisions ranging from 87 % to 98 % and recalls from 59 % to 94 %.

Index Terms— video mining, video structuring, multimodality, mutual information, clustering, content extraction

1. INTRODUCTION

Video structuring is a crucial step in most video content analysis tasks with applications in archiving, browsing or re-purposing. Generally speaking, the structuring task consists in finding time stamps of events related to the underlying structure of an audiovisual material. From this general definition of structuring, two variants can be distinguished: (i) dense segmentation, where structural events explain the entire video, and (ii) detection of specific events relevant to the semantic or editing structure of the video, such as anchor shots, transition animated screens, advertisements, or goals in sport videos.

Many work have addressed event based video structuring using supervised approaches for which models—be they rule based, probabilistic or discriminative—are trained from manually annotated data. This supervised approach has been extensively used for anchor person detection [1, 2, 3] or event detection in sport videos [4, 5]. However, such methods are poorly generic and lack robustness across genres and channels. Models have to be trained for each event considered or whenever an event changes appearance. Taking the example of anchor person shots, most approaches are highly specific of a particular channel or program and fail to generalize. As a mean towards generic and robust structural event detection, we propose a video mining technique to discover audiovisually consistent repeating segments among which those likely to be relevant to the editing structure of the video are selected.

Mining videos for repeating patterns has received attention over the past decade. Several approaches have focused on the discovery of near-duplicate repetitions [6, 7, 8] but cannot deal with the crucial issue of variability across repetitions. Since structurally relevant events are often characterized by their strong visual consistency, the problem of mining repeating structural elements has been addressed using clustering techniques (see, e.g., [9, 10, 11, 12, 13, 14]). But several problems arise from clustering, such as deciding the optimal number of clusters or dealing with outliers. This last point is particularly crucial since, in most cases, only those shots relevant to the structure exhibit similarity and should therefore be part of a cluster. Finally, models can be exploited to jointly discover structural events and the corresponding models [15, 9, 16] but these approaches assume dense segmentation and cannot address the discovery of sporadic structural events.

This short survey of the literature points out two striking facts. Firstly, multimodality, in particular relations between modalities, is seldom exploited to discover events of interest and one rather assumes either visual or audio coherence. Secondly, apart from the dense segmentation case, most discovery techniques do not exploit models, in particular discriminative ones, which have proven highly efficient for event detection.

In this paper, a multimodal event mining technique is proposed to discover repeating video segments exhibiting audio and visual consistency. The mining strategy first exploits independent audio and visual cluster analysis to provide segments which are consistent in both their visual modality and their audio modality, thus likely corresponding to a unique underlying event. A subsequent modeling stage using discriminative models enables accurate detection of the underlying event throughout the video, automatically selecting positive and negative training examples from the output of the initial mining stage. This event discovery technique is applied to unsupervised video structuring, using simple heuristics on the occurrence patterns to select those audiovisually consistent events which are relevant with respect to the editing structure of the video.

The paper is organized as follows. The event mining stage is detailed in Section 2. Section 3 discusses selection of events relevant to the structure. Experimental setup and results are provided in Section 4 before concluding.

2. CROSS-MODAL DISCOVERY OF AUDIOVISUALLY CONSISTENT EVENTS

The cross-modal mining algorithm at the core of this work aims at discovering repeating events in the video with high audiovisual consistency across occurrences. The reason for this choice is that events with such characteristics are usually related to higher level information like the editing structure of a TV program or stream, which in turn often indicates a change in semantic content. The proposed method can discover multiple low-level audiovisual events among

978-1-61284-350-6/11/$26.00

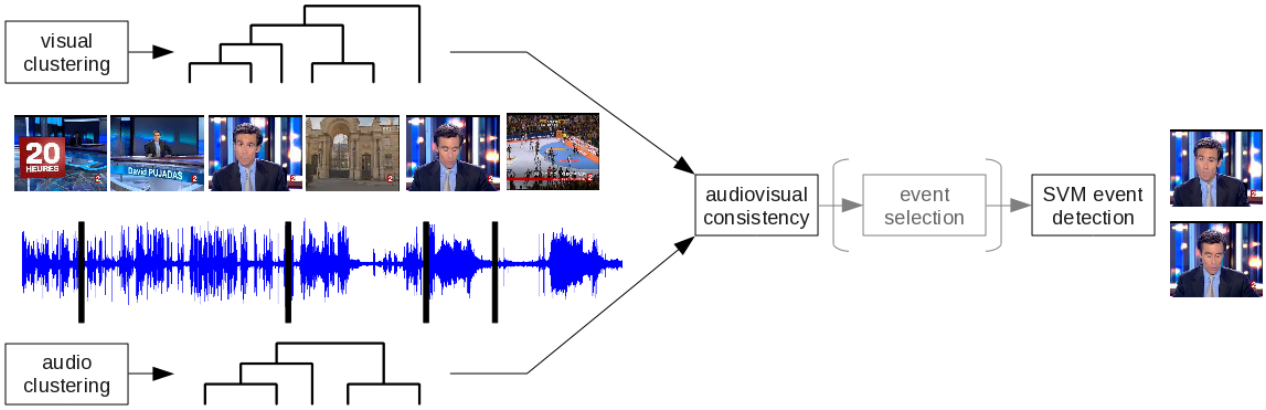

Fig. 1. General principle of the proposed cross-modal discovery process. Note that the figure includes the selection of relevant events (in grey), which is not per se part of the mining process.

which the ones relevant to the structure will be selected based on simple rules as described in Section 3.

In this section, an overview of the method is first given before detailing the key components, namely, mining of candidate events and detection of those events using support vector machines (SVM).

2.1. Overview of the discovery method

Figure 1 schematically illustrates the various steps of the overall discovery method. Firstly, the soundtrack is segmented into homogeneous audio chunks. Similarly, the visual track is segmented into shots. For each modality, bottom-up clustering provides a hierarchical structure of nested clusters, known as dendogram, where each node in the dendogram represents a set of segments. To mine audiovisually consistent events, cross-modal consistency is measured between each pair consisting of one node from the audio dendogram and one node from the visual dendogram, the most consistent pairs defining candidate events. Note that this approach significantly differs from typical ones where the clusters of interest are selected from a fixed, unique clustering, be it monomodal or multimodal, using intra-clustering measures or heuristics. Finally, each candidate event is accurately detected using a SVM-based discriminative classifier trained in an unsupervised manner. To learn this classifier, positive and negative training samples are automatically selected from the results of the mining step, further exploiting cross-modality information. Note that, though not part of the mining algorithm itself, selection of candidate events relevant to the structure (grey box in Figure 1) might be performed between discovery and SVM detection for efficiency reasons that will be discussed in the Section 3.

2.2. Discovery of candidate events

Segmentation and clustering, independently performed in the audio and visual modalities, provide two dendograms where each node is a set of segments in the corresponding modality. Finding out a repeating event with high similarity in both its audio and visual content relies on a measure of the audiovisual consistency between a cluster of audio segments and a cluster of visual shots. Audiovisual consistency is measured using mutual information, a measure of the information shared by two random variables. Clearly, two inconsistent clusters should share little information.

Let us denote C_i^x the i-th cluster for the modality $x \in \{\mathcal{A}, \mathcal{V}\}$, as defined by the i-th node of the corresponding dendogram. The mutual information between $C_i^{\mathcal{A}}$ and $C_j^{\mathcal{V}}$ is defined as

$$I(C_i^{\mathcal{A}}, C_j^{\mathcal{V}}) = \sum_{a\in\{0,1\}} \sum_{v\in\{0,1\}} p(a,v) \ln \left(\frac{p(a,v)}{p(a)p(v)} \right) \quad (1)$$

where a and v are binary random variables indicating membership to $C_i^{\mathcal{A}}$ and $C_j^{\mathcal{V}}$ respectively. Practically the probabilities $p(a,v)$, $p(a)$ and $p(v)$ are estimated from the temporal segmentations. For example, the joint probability $p(a=1, v=1)$ is estimated as the amount of time jointly corresponding to a segment of $C_i^{\mathcal{A}}$ and one of $C_j^{\mathcal{V}}$ (dark grey area in Figure2), normalized by the total duration of the video. In other words, Eq. 1 indicates to what extent the segments in the audio cluster $C_i^{\mathcal{A}}$ coincides with those of the shot cluster $C_j^{\mathcal{V}}$. Moreover, since clusters are selected from the dendograms, they exhibit a certain amount of homogeneity. These two facts guarantee that two clusters with high mutual information define audiovisually coherent segments.

The use of a cross-modal information measure has several invaluable advantages over conventional clustering methods. First of all, the problem of determining the best cluster—in our case, the one with the highest audiovisual coherence—is skirted. Indeed, conventional clustering approaches rely on somewhat arbitrary criteria to decide where to stop clustering. Moreover, among the resulting clusters, one has to find out the ones that best correspond to the application needs. By measuring the mutual information between all pairs of audio and visual clusters, these two problems are avoided. Another advantage of the cross-modal approach is that all possible pairs of clusters can be ranked from the more consistent to the less, thus providing multiple candidate events where conventional approaches usually output a single one. Having multiple solutions makes it possible to discover several characteristic events, for example, each guest in a talk show, and to filter out irrelevant ones using heuristic rules such as the one discussed in Section 3.

Note that, as the method explained above for the discovery of candidate characteristic events is fully based on a cross-modal measure of consistency, the work presented in this paper does not cover the case of audio-only or video-only repeats in the video. However such monomodal events could be searched for in a second step using appropriate techniques on a substantially reduced amount of data.

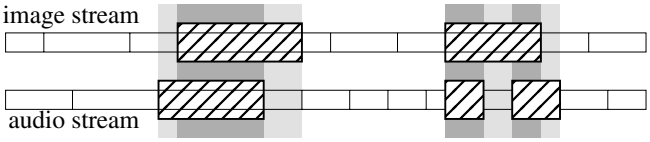

Fig. 2. Illustration of the combination of the audio and shot segmentations. Squared elements with diagonal filling denote, for each modality, the segments corresponding to the event selected. Dark grey regions define the intersection of the two segmentations where the event is present in both modalities, while light grey ones define the union where the event is present in either one of the modalities.

2.3. Unsupervised discriminative modeling

After the initial discovery step, a candidate event is characterized by a pair $\{C_i^{\mathcal{A}}, C_j^{\mathcal{V}}\}$, with the corresponding temporal segmentations $\{S_i^{\mathcal{A}}, S_j^{\mathcal{V}}\}$. Note that since segmentation and clustering are performed independently for each modality, $S_i^{\mathcal{A}}$ and $S_j^{\mathcal{V}}$ differs and must be combined to define the extent of the event under consideration, as illustrated in Figure 2. Early experiments have shown that taking segments occurring jointly in $S_i^{\mathcal{A}}$ and $S_j^{\mathcal{V}}$ (i.e., taking the intersection of the two segmentations) results in a high precision for the candidate event, however with poor recall. Oppositely, taking the union of the two segmentations increases recall while decreasing precision. The properties of the intersection and union of the two segmentations can be exploited to select positive and negative training samples in a totally unsupervised manner, so as to enable the use of a discriminative classifier in a final detection step. Following the idea that the intersection of the audio and video segmentations gives high precision, while their union gives high recall, the training samples were selected as follows.

First, an audiovisual segmentation of the video is constructed by merging the boundaries of the audio and visual segmentations. Each resulting segment, s_k, is described by an audiovisual feature vector $\vec{AV}_{s_k}$ which consists of the concatenation of the two feature vectors $\vec{A}_{s_l}$ and $\vec{V}_{s_m}$, describing respectively the current audio segment s_l and the current video shot s_m. Then, for an event characterized by the two clusters $\{C_i^{\mathcal{A}}, C_j^{\mathcal{V}}\}$ and the corresponding segmentations $\{S_i^{\mathcal{A}}, S_j^{\mathcal{V}}\}$, the positive ($+\mathbb{1}$ class) and negative ($-\mathbb{1}$ class) training samples are selected according to

$$\vec{AV}_{s_k} \in +\mathbb{1} \quad \text{if} \quad s_k \subset S_i^{\mathcal{A}} \cap S_j^{\mathcal{V}}$$
$$\vec{AV}_{s_k} \in -\mathbb{1} \quad \text{if} \quad s_k \not\subset S_i^{\mathcal{A}} \cup S_j^{\mathcal{V}}$$

In other words, vectors corresponding to the intersecting parts of $S_i^{\mathcal{A}}$ and $S_j^{\mathcal{V}}$—a region where we expect low false positives—are chosen as positive training samples, while all those corresponding neither to $S_i^{\mathcal{A}}$ nor $S_j^{\mathcal{V}}$—a region where we expect low false negatives—are selected as negative training samples. Here again, cross-modal information (i.e., A/V intersection and union segmentations) is fully exploited as a means to automatically select pseudo training samples from the originally unlabeled data.

Obviously, unsupervised selection of training samples cannot ensure that the training data is error free. Hence, training data are bound to contain false positive and false negative examples, which can be seen as outliers for their class. Support vector machines, known for their robustness to outliers, were thus used to model the data using a RBF kernel. Hyper-parameters were optimized in a unsupervised manner through a grid search with a 5-fold cross-validation procedure on the training set. Finally, given the SVM trained with the optimal hyper-parameters, all segments in the video are classified as either corresponding to the event considered or not, thus terminating the entire discovery algorithm.

3. SELECTING RELEVANT EVENTS FOR STRUCTURE ANALYSIS

The discovery process finds audiovisually coherent events regardless of their relevance to some semantic information need. In the context of a particular application, re-ranking is needed to emphasize events of interest in this context. In the case of video structure analysis, not all audiovisually coherent events are relevant. For example, in a news show ending with a guest's interview, it was observed that the mining step ranked shots of the guest speaking as the most coherent event instead of the expected anchor shots which are much more relevant to the structure of the program. Thus minimal expert knowledge about the targeted event topology needs to be incorporated in the system at this point. One property of structural events is that they occur more or less regularly throughout the video and therefore spans the entire program. Based on this property, structurally relevant events should meet the following requirements in addition to audiovisual consistency: (1) span a large part of the program, (2) exhibit time regularity both in their duration and in their frequency, and (3) be sparse.

Heuristic filtering rules are used to ensure these properties and to avoid partial discovery of events. Candidate events that do not span at least half of the program's duration as well as those with a single occurrence are discarded to fulfill requirement 1. Re-ranking the remaining events is done according to

$$S(k) = \frac{m_s(k)}{\sigma_s(k)} \frac{m_i(k)}{\sigma_i(k)} \tag{2}$$

where $m_s(k)$ and $\sigma_s(k)$ are respectively the mean and standard deviation of the duration of the (discovered) occurrences of the k-th event, and $m_i(k)$ and $\sigma_i(k)$ are the mean and standard deviation of the duration of the intervals between occurrences. The denominator term, related to requirement 2, benefits to solutions with high regularity in their appearance patterns. In the numerator, $m_i(k)$ enables sparsity while $m_s(k)$ aims at finding out the longest possible event, hence avoiding event splitting (i.e., an actual event whose occurrences are spread across two or more discovered segments).

Note that event selection should ideally be performed at the end of the entire mining process. However, to avoid unnecessary and computationally demanding SVM training and classification steps, event selection was performed after the initial discovery step, considering only the N most consistent events. After removing irrelevant events from the N-best list of candidates, the remaining ones were sorted according to Eq. 2 and, in this work, only the best resulting event was further processed using SVM classification.

4. EXPERIMENTAL SETUP AND RESULTS

Experiments were carried out on a comprehensive data set including various types of TV shows. As no publicly available data set exists for such task, TV programs were recorded from various French television channels and segments corresponding to structural events were annotated by a human expert. We first describe the data and briefly present the video and audio segmentation and clustering algorithms used before discussing evaluation criteria and presenting results.

Fig. 3. Typical structural events. From left to right and up to down: anchor person in news, 2 flash news screens for 2 different programs, magazine anchor person, guest in a talk show, contestant presentation in a game.

Type	#shows	#occs	duration
Flash news	2	4	0h40
News	1	2	1h20
Magazine	2	4	3h
Investigation	3	7	6h
Talk show	2	3	3h
Games	1	2	0h40
Total	11	22	14h40

Table 1. Description of the data set: #shows is the number of shows with a different name; #occs is the total number of occurrences; duration is the total duration over all occurrences.

4.1. Data set

Discovering structural events through audiovisual consistency in videos only makes sense for program genres in which such events are present. Evaluation is therefore performed on a selection of programs from the following genres: flash news, news, magazines, investigation reports, talk shows and games. The data set consists of a few shows for each genre, collected over various French television channels, as detailed in Table 1. A brief description of each genre and of the related key structural elements is provided below, key frames of typical events of interest being illustrated in Figure 3.

News denote the classical news show where appearances of the anchor speaker(s) constitute the main structural event(s). Flash news are short news programs where no anchor speaker appears on screen. Rather, news items are separated by a typical computer-generated screen displaying the title of the next item with a jingle as background music. The variability between occurrences of the anchor event is less than for anchor person shots and mainly comes from the title of the news item.

Magazines denote news-like shows tackling non-news subjects. They are less formal than classical news shows in their structure, resulting in an increased variability of the structurally relevant events, and are therefore more challenging for mining purposes. Two different magazine shows are present in the data set: one consists of reports or interviews separated by sequences showing one or two anchor persons with various set-up (an example is shown in the leftmost column, bottom row of Figure 3) while the other one consists of journalists presenting sport news in turn, each news item being separated by a computer-generated screen.

Investigation reports are programs where a few topics are investigated in-depth. As for news and magazines, they are presented by one or two anchor persons. In addition to variability, the challenge lies in the fact that anchor shots occur only a few times over a long period. Data were collected from three different programs with duration ranging from 1 hour to 2 hours, the longest one exhibiting two anchor persons.

Finally, talk shows and games significantly differs from news-like contents and are much more challenging. In particular, such programs often exhibit several structuring events from an audiovisual standpoint. For example, in talk shows, a close-up of a particular participant is a structuring event that occurs several times across the program, roughly every time the participant is speaking.

4.2. Segmentation and clustering

Classical segmentation and clustering techniques were used to provide independent audio and visual dendograms.

Segmentation of the audio stream was performed using the Bayesian information criterion (BIC). Bottom-up agglomerative clustering is based on a variant of the BIC criterion to select the two closest clusters, where clusters are modeled using 16 component Gaussian mixture models [17]. Each audio cluster is characterized by a unique 528-dimensional feature vector gathering scaled versions of the 33-dimension mean vectors of the 16 Gaussian components.

The video stream is segmented into shots based on color histograms to detect abrupt changes and progressive transitions. Each of the resulting shot is summarized by a key frame, taken in the middle of the shot, in turn represented as a RGB histogram with 8 bins per color. Bottom-up clustering relies on the Euclidean distance between the 512-dimension color histograms using Ward's linkage.

The input samples of the SVM-based event detection module (see section 2.3) are thus 1040-dimensional audiovisual feature vectors, each resulting from the concatenation of the two feature vectors describing respectively the current audio segment and the current video shot.

4.3. Evaluation criteria

Evaluating mining algorithms raises two main issues: mapping the elements discovered to semantic events and evaluating the quality of the discovery job in regard of the mapping.

By construction, the mining process aims at finding out events with common low-level characteristics—in our case a strong audiovisual consistency—, in the hope that they correspond to some semantic event. Regarding editing structure analysis as the targeted application, ground truth for evaluation was constructed by systematically annotating all low-level events in the videos considered relevant from a structural viewpoint. Each event discovered is mapped to the best matching structural event in the ground truth to compute performance statistics. Note that, though discovering multiple events is possible with the proposed method, we did not considered this case and evaluation is limited to the discovery of a single event per program for sake of simplicity.

Given the occurrences of a discovered event and the occurrences of the best corresponding semantic event as annotated in the video, the quality of the discovery process is classically measured in terms of precision (P), recall (R) and F_1 measure. Precision corresponds to the proportion of discovered events actually corresponding to the mapped semantic event. Recall corresponds to the proportion of the semantic event actually included in the output of the discovery process. Precision and recall are computed on a time basis—as opposed

	Mining + selection			+ SVM detection		
Genre	R	P	F_1	R	P	F_1
Flash News	0.57	**1**	0.71	**0.76**	0.98	**0.85**
News	0.85	**1**	0.91	**0.94**	0.95	**0.94**
Magazine	0.63	**0.91**	0.73	**0.76**	0.89	**0.81**
Investigation	0.47	0.95	0.59	**0.59**	**0.96**	**0.69**
Talk show	0.51	**0.88**	0.64	**0.76**	0.87	**0.81**
Games	0.72	**0.95**	**0.8**	**0.75**	0.92	**0.8**

Table 2. Structural event detection performances

to an event basis—and, for each genre, performance measures are averaged across occurrences.

4.4. Results

Results are reported separately for each type of programs as defined in Section 4.1. Moreover, to illustrate the respective contribution of each step of the algorithm, results are reported for the structural audiovisual event discovery stage alone (mining+selection), and after SVM detection. For the sole discovery stage, segments for an event are defined as the intersection between the corresponding audio and video temporal segmentations (see discussion in section 2.3). All results are summarized in Table 2.

4.4.1. Qualitative analysis

Detailed analysis of the results has revealed that, when applying selection after the mining step, the best events selected were actually relevant from a structural viewpoint for all programs in the data set. This is not always the case when no selection is performed, thus proving the effectiveness of the requirements fixed to filter out irrelevant solutions for structure analysis purposes.

For all flash news program, the computer-generated screens with an audio jingle separating the news items (middle and right-most samples in the top row of Figure 3) were systematically identified as the most structuring events. This was also the case for the sports magazine program. In the other magazine program where two anchor persons appear, shots showing the male anchor in a double screen set up (left-most sample in the bottom row of Figure 3) were returned by the algorithm. This shots actually are very relevant for the editing structure of the program as they correspond to the introduction of the coming report. For news and investigation programs, the most relevant events discovered correspond in all cases to occurrences of anchor person shots though sometimes partially retrieved. In the two investigation programs with two anchor persons, the algorithm focused on either one of the anchors, only returning occurrences of him/her speaking. Similarly, for the talk show program, the algorithm focused on appearances of one of the main guest speaking all along the program. Finally, for the game programs, the algorithm returned screens with a particular common visual set up and the same off-voice either presenting the participants, the coming game section or the gift given to the loosers at the end of a section.

Regarding computation time, programs were processed in half real-time (on a recent laptop), most of the computation (about 80 %) time being used for clustering audio and video segments.

4.4.2. Discovery step

In the 3 leftmost columns, results are reported after selecting the best event from the N-best list provided by the initial discovery step using the heuristic described in Section 3. Clearly, the number N of events retained from the discovery step results in a trade-off between the guarantee to select an audiovisually consistent event and the appropriateness of the solution with respect to the properties of structural elements. Too short a list guarantees audiovisual consistency but may not contain structuring events. On the contrary, a longer list might result in the selection of an event with good structuring properties but lacking audiovisual consistency. In these experiments, the list size was experimentally fixed to $N = 5$ which globally gave the most satisfying results across all programs. Obviously, the optimal value of N might be dependent upon the program's type. However, to keep the algorithm generic and keep the amount of supervision to a minimum, N was fixed across all programs.

Results in Table 2 clearly confirm that the "intersection" solution indeed lead to discovering occurrences of structural events with a high precision. Across the various program types, precision ranges from about 90 % to 100 %. Recall, however, usually remains insufficient, in particular for investigation reports and talk shows for which 50 % is hardly achieved. The final SVM-based detection step aims at compensating this drawback.

4.4.3. Detection step

Results after the SVM detection step[1] in the rightmost columns of Table 2, show systematic significant gains in recall with only minimal decrease in precision with respect to the discovery step. Recall values ranging from 47 % to 85 % after the discovery step are increased between 59 % and 94 % after unsupervised SVM classification. In particular performances are near perfect for the news programs, with a recall of 94 % and a precision of 95 % for the detection of close-up anchor shots. Exceptions are the game programs for which no improvement is obtained on the F_1 measure by the SVM-based detection step. These results validate the unsupervised selection of training data from the discovery step and shows that discriminative classifiers can be used in a totally unsupervised context in order to enhance the retrieving capacity of a discovery system, as long as an effective method for selecting training samples is used.

In spite of good results on several program types, progress are still needed to improve recall, in particular in difficult cases like the investigation programs where the audiovisual set-up may vary significantly.

5. CONCLUSION

A multimodal strategy for unsupervised mining of videos from audiovisually consistent segments was presented and applied to the discovery of events relevant to video structuring tasks. The use of multiple modalities to discover characteristic events enables to circumvent problems related to clustering in a single modality, in particular selecting the adequate stopping criterion and cluster. Rather, optimal clustering is found by comparing multiple audio and visual clusters as given by independent bottom-up hierarchical clustering in each modality. Experimental results show that this cross-modal mining step, associated to event selection, is efficient, yielding high precision however with insufficient recall. Recall is improved using discriminative models whose parameters are trained in a totally

[1]The publicly available libsvm implementation was used (C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001).

unsupervised manner, automatically selecting training samples from the initial mining and selection step.

From a more general standpoint, this work demonstrates the feasibility of unsupervised discovery of significant events for semantic tasks such as structure analysis. This opens numerous opportunities in the field of multimedia mining and content-based analysis. Indeed, the event discovery method discussed in the paper is in fact quite generic and not restricted to structure analysis. Mining is based on the general principle that events of interest share at least two common characteristics. In the framework of structure analysis, we limited ourselves to the two following characteristics: visual consistency—via color histograms—and audio consistency—via speaker and sound class modeling. But any other characteristic could be used instead of, or in addition to, either one. For example, movement based clustering seems interesting for event discovery in sports video, while local visual interest points may probably allow clustering of shots with similar objects.

Finally, it was mentioned that the proposed mining method is not limited to the discovery of a single event and can provide a ranked list of characteristic events. We believe that this is a particularly interesting feature of the proposed cross-modal algorithm that has applications in many domains. Indeed, apart from news and flash news, most programs exhibit more than one structurally relevant event. Many news programs have two anchor persons. Magazines and talk shows often include guests which repeatedly appear throughout the show. Discovering multiple events is then necessary to reveal the structure of such programs. However, providing a ranked list of the N-best events is only a first step towards structure analysis from multiple events. In particular, selecting the relevant events among the N bests remains an open issue. How to ensure that there is no conflict in the discovery process, for example with SVM classification, is another one. Finally, accounting for multiple events in the structure probably requires expert knowledge on characteristic patterns of structural events.

6. ACKNOWLEDGMENT

This work was partly funded by OSEO, French State agency for innovation, in the framework of the Quaero research program.

Many thanks to Mónica Corlay for doing the ground truth annotation work, and to Sébastien Campion for providing Python code for key frame clustering, as part of the PimPy library[2].

7. REFERENCES

[1] H. J. Zhang, C. Y. Low, S. W. Smoliar, and J. H. Wu, "Video parsing, retrieval and browsing: an integrated and content-based solution," in *ACM Intl. Conf. on Multimedia*, 1995, pp. 15–24.

[2] A. Hauptmann, R. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H. Wactla, "Informedia at TRECVID 2003: Analyzing and searching broadcast news video," in *Text Retrieval Conference*, 2003.

[3] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *ACM Intl. Conf. on Multimedia*, 2004, pp. 656–659.

[4] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, "Multi-modal extraction of highlights from TV Formula 1 programs," in *IEEE Intl. Conf. on Multimedia & Expo*, 2002, pp. 817–820.

[5] F. Wang, Y.-F. Ma, H.-J. Zhang, and J.-T. Li, "A generic framework for semantic sports video analysis using dynamic bayesian networks," in *Intl. Multimedia Modelling Conference*, 2005, pp. 115–122.

[6] M. Covell, S. Baluja, and M. Fink, "Detecting ads in video streams using acoustic and visual cues," *IEEE Computer Magazine*, vol. 39, no. 12, pp. 135–137, Dec. 2006.

[7] C. Herley, "ARGOS: Automatically extracting repeating objects from multimedia streams," *IEEE Trans. on Multimedia*, vol. 8, no. 1, pp. 115–129, Feb. 2006.

[8] A. Jacobs, "Using self-similarity matrices for structure mining on news video," in *Advances in Artificial Intelligence – Helenic Conf. on Artificial Inteligence*, 2006, vol. 3955 of *LNAI*, pp. 87–94.

[9] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 1999, vol. 6, pp. 3037–3040.

[10] A. Divakaran, K. Peker, R. Radharkishnan, Z. Xiong, and R. Cabasson, "Video summarization using mpeg-7 motion activity and audio descriptors," in *Video Mining*, D. DeMenthon A. Rosenfeld, D. Doermann, Ed. Kluwer Academic, Oct. 2003.

[11] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchroperson detection for news video story parsing," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 9, pp. 765–775, 2002.

[12] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori, "Unsupervised discovery of action classes," in *IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1654–1661.

[13] C. Ma and C.-H. Lee, "Unsupervised anchor shot detection using multi-modal spectral clustering," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 2008, pp. 813–816.

[14] M. Broilo, E. Zavesky, A. Basso, and F. De Natale, "Unsupervised event segmentation of news content with multimodal cues," in *Intl. Workshop on Automated Information Extraction in Media Production*. 2010, pp. 39–44, ACM.

[15] M. Naphade, C. Li, and T. Huang, "Discovering recurrent events in multichannel data streams using unsupervised methods," in *Data Mining – Next generation challenges and future directions*, pp. 147–156. AAAI Press, Oct. 2004.

[16] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Unsupervised mining of statistical temporal structures in video," in *Video Mining*, A. Rosenfeld, D. Doremann, and D. DeMenthon, Eds., chapter 10. Kluwer Academic Publishers, 2003.

[17] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms," in *Intl. Conf. on Speech and Language Processing*, 2004.

[2] http://pim.gforge.inria.fr/pimpy/